新疆融媒体报道热点领域提取与分析

——基于天山网新闻标题语料库的计量研究

宫 媛 冯玮杰
（新疆大学，新疆 乌鲁木齐 830046）

摘 要：本文运用手工录入与 Python 相结合的方法，收集了天山网新疆新闻标题，建立"天山网新闻标题语料库"，并分别使用 NLPIR-ICTCLAS 汉语分词系统和 MyZiciFreq 字词频率统计工具对语料进行分词处理和词频统计。本文对天山网新疆新闻标题的统计和社会价值分析，为新闻标题的编写提供了参考依据，同时也展现了新疆地区融媒体所关心的热点事件，也体现出了新疆各族人民团结一心的坚定决心。

关键词：新闻标题；语料库；词频；天山网；融媒体 **中图分类号：**H087 **文献标识码：** A
文章编号：1671-0134（2022）04-036-04 DOI：10.19483/j.cnki.11-4653/n.2022.04.008
本文著录格式：宫媛，冯玮杰．新疆融媒体报道热点领域提取与分析——基于天山网新闻标题语料库的计量研究 [J]. 中国传媒科技，2022（04）：36-38，64.

导语

新闻是传播要点信息、记录社会事件以及反映时代风貌的一种文体。随着社会信息化程度的加深，新闻步入了"融媒体"时代。但语言文字仍然是新闻传播的重要媒介。新闻标题是新闻的重要组成部分，新颖有趣的新闻标题可以吸引读者阅读新闻正文。而新闻标题中的一些关键高频词可以反映出某一时间段的社会关注点。因此，新闻标题语言的词频研究对于新闻标题的制作、反映社会面貌具有一定意义。

关于新闻标题语言的词频研究，马子恩（2012）《热点事件新闻语料库的研制及词汇研究》一文以《扬子晚报》部分热点事件新闻标题作为原始语料，研究新闻语域词频分级、热点事件特殊词汇的分类。[1] 班文冲（2016）《基于语料库的网络新闻标题词频研究——以人民网、新华网和新浪网、网易网为例》，在对处理后的原始语料的基础上制成 4 个语料库，并对高频词统计进行了研究。[2]

但是，鲜有学者对新疆新闻进行研究，所以，本文选取天山网新疆新闻的标题，建立"天山网新闻标题语料库"并对新闻标题进行分析研究。对新闻标题的词频分析，可以直接从统计的角度分析高频词的使用情况、词频与词频排名的关系，这些分析可以直观地看出词频与其他要素间的关系，有助于新闻编写者在编写新闻时挑选合适的词语。对于新闻标题的社会价值分析，有助于读者理解该时间段在该地区的热词及所发生的热点事件。

1. 新闻标题语料库来源及处理

1.1 新闻标题语料库来源

语料库（Corpus）是指经过科学取样和加工的大规模电子文本库。[3] 借助计算机语言分析工具后，研究者可以开展相关的语言理论与应用研究。语料库语言学研究的基础就是语料库，它被广泛地应用于语言教学、自然语言处理（NLP）等方面。

本文选用天山网新闻标题来进行高频词统计研究。天山网是新疆维吾尔自治区唯一一家重点新闻宣传网站，由新疆维吾尔自治区党委宣传部、人民日报网络中心合作建设，由新疆维吾尔自治区人民政府新闻办公室主管、新疆新媒体中心承办。

1.2 新闻标题语料库获取方式

本文采用手工录入的方法，收集了 2019 年 1 月 29 日至 2020 年 11 月 4 日天山网新疆新闻标题，共计 3659 条标题作为语料源，建立了新闻标题语料库，这些新闻标题涉及政治、经济、文化等各个方面。

例 1. 陈全国：坚定不移推动党中央各项改革部署落到实处（2019 年 2 月 15 日，时政类）

例 2. 去年新疆口岸与"一带一路"沿线国家进出口额同比增 13.5%（2019 年 2 月 15 日，经济类）

例 3. 额敏县：新春活动加保健知识宣传 超赞（2019 年 2 月 20 日，文旅类）

1.3 语料的分词处理

对中文的自然语言处理，分词是基本的预处理手段之一。[4] 只有先把原始语料以空格为分界符，分出一个个最小的能够独立运用的词，才可以进行词频统计。不然，只能对原始语料进行字频统计，而不是词频。所以，本文进行的词频研究，必须在对原始语料进行中文分词后进行研究。

本文使用的是 NLPIR-ICTCLAS 汉语分词系统。该软

基金项目：本文系国家社会科学基金一般项目"基于语料库统计的新疆全媒体新闻语言研究"（项目编号：20BYY072）的研究成果。

件由北京理工大学的张华平教授领导的大数据挖掘与搜索实验室研发，具有一定的权威性与准确性。例如：

例 4. 元宵节出疆机票价格有折扣（2019 年 2 月 15 日）

元宵节 出 疆 机票 价格 有 折扣

从以上新闻标题的分词例子中可以看出，NLPIR-ICTCLAS 汉语分词系统基本可以准确切分出汉语词，所以该软件可以满足本文的研究要求。

2. 新闻标题的词频统计

在对原始语料进行分词处理后，本文对分词后的语料进行词频统计，得出某一个词在语料库中的词频、占比以及排名。本文使用国家语言文字工作委员会开发的 MyZiciFreq 字词频率统计工具进行词频统计，具有权威性和科学性。该软件可以自动对分词后的语料进行词频统计，并输出词频和占比。

天山网新疆新闻标题词表共计有 73951 个字，包含 6424 个词条，总词次为 42381。在第一列中将该词表中的词频以 1、2、3……进行排序，在第二列中列出相对应排名的词汇，在第三列中生成词频，词频统计是全面统计该词在语料库中出现的总次数，在第四列中计算出该词在语料库中的占比，计算方法是：占比 = 词频 / 总词次，在得出占比数据之后把小数保留至小数点后两位，以保证精确性，并把得出的占比数据依次生成到第四列中。

这样直接生成的高频词表，包含了大量无特色的词条，例如“月”“日”“的”“大”“是”等。所以，将部分无意义、无特色的高频词剔除，并从天山网新疆新闻标题词表中，按照词频排名的顺序，选出具有社会性、地域性的词条进行递补，得出天山网新疆新闻标题词表最高频的 30 个有特色的词条（见表 1）。并按高频词所代表的领域，把它们归为四大类（见表 2）。

表 1　天山网新疆新闻标题词表具有特色的词频最高的 30 个词

排名	词	词频	占比	排名	词	词频	占比
1	新疆	1529	3.61%	16	天气	78	0.18%
2	乌鲁木齐	688	1.62%	17	工作	77	0.18%
3	新增	199	0.47%	18	大部	76	0.18%
4	自治区	191	0.45%	19	铁路	75	0.18%
5	兵团	165	0.39%	20	中国	71	0.17%
6	病例	137	0.32%	21	高温	69	0.16%
7	确诊	125	0.29%	22	气温	65	0.15%
8	肺炎	108	0.25%	23	交警	63	0.15%
9	疫情	91	0.21%	24	高速	55	0.13%
10	企业	89	0.21%	25	脱贫	54	0.13%
11	旅游	82	0.19%	26	旅客	53	0.13%
12	项目	82	0.19%	27	国际	52	0.12%
13	服务	79	0.19%	28	机场	51	0.12%
14	建设	79	0.19%	29	就业	42	0.10%
15	启动	78	0.18%	30	景区	41	0.10%

表 2　高频词分类

排名	高频词类	词	词数	总词频
1	地名类	新疆 乌鲁木齐 自治区 兵团 中国	5	2644
2	文旅类	旅游 服务 天气 大部 铁路 高温 气温 交警 高速 旅客 国际 机场 景区	13	839
3	防疫类	新增 病例 确诊 肺炎 疫情	5	660
4	发展类	企业 项目 建设 启动 工作 脱贫 就业	13	501

在表 2 中，高频词被分为了防疫类、地名类、文旅类和发展类四大类。除去地名类仅代表新闻报道的地区，不具备体现新闻热点的特性。可以发现，在 2019 年 1 月 29 日至 2020 年 11 月 4 日这个时间段内，新疆融媒体的新闻报道把重心放在了防疫、文旅和发展上。透过高频词来看的话，疫情、文旅以及发展就是新疆在 2019 年 1 月 29 日至 2020 年 11 月 4 日内的热点事件。下面，依据表 1 和表 2，进一步讨论每一个高频词大类中的词出现的新闻标题，以及从社会性、地域性的角度，对大部分高频词进行分析。

第一，由于该表的语料来源是天山网新疆新闻，所以在高频词中出现了很多具有地域性的词条。例如，“新疆”“乌鲁木齐”“自治区”“兵团”等词条。包含这些词条的新闻标题，基本上涵盖了新疆的各个方面，对新疆的宣传工作作出了很大的贡献。例如，

例 5. 新疆喀纳斯湖畔迎来全国 700 对佳人 还创造一项为爱表白的世界纪录（2019 年 7 月 3 日）

例 6. 乌鲁木齐市米东区千人同吃 1.7 米巨碗“国庆面”（2019 年 10 月 1 日）

例 7.2018 年生产总值比上年增长 6% 兵团粮棉产量呈现“双增长”（2019 年 2 月 15 日）

所以，包含这些高频词条的新闻标题反映了新疆发展稳中带好，呈现出了社会繁荣稳定、人民安居乐业的良好局面，充分展现了党总揽全局、协调各方的领导核心作用。同时，在地名类里，还有一个高频词是“中国”，例如：

例 8. 想借涉疆“法案”干涉牵制中国，只能是痴心妄想（2019 年 12 月 6 日）

“中国”一词的高频出现，表现了新疆融媒体不仅立足于新疆本土而且放眼全国的广阔视野。更加体现了，在中国共产党和中国政府的坚强领导下，随着“两个一百年”奋斗目标和中华民族伟大复兴中国梦的实现，新疆会奋力书写好中国特色社会主义的新疆篇章。

第二，新疆的旅游业很发达，新疆一直是全国各族人民旅游的首选地之一。所以，在该表中，高频词也包括了许多与旅游、交通和天气相关的词条，例如，“旅游”“天气”“铁路”“高温”“气温”“交警”“高速”“旅客”“国际”“机场”“景区”等词条。这表现了新疆坚持绿色发展，

努力建设天蓝地绿水清的美丽新疆。例如：

例 9. 富蕴县加快景区建设推动旅游业高质量发展（2019 年 4 月 10 日）

例 10. 乌鲁木齐市今日天气晴好 最高气温 30℃（2019 年 6 月 19 日）

例 11. 新疆升级 53 对高速公路服务区让旅游更畅行（2019 年 7 月 12 日）

例 12. 从新疆国际大巴扎到喀纳斯：国庆假期游客畅游新疆好地方（2019 年 10 月 7 日）

这些与旅游相关的高频词条，从侧面体现了自“旅游兴疆”战略实施以来，新疆各地区以旅破题，以旅游为龙头、带动产业发展的格局初步形成，新疆新闻业也通过旅游宣传的工作，向疆外持续展示“大美新疆”的形象。要充分认识实施旅游兴疆战略的重大意义，切实把大力发展旅游业摆在关系各族人民福祉、关系社会稳定和长治久安的战略高度，推动旅游业高质量发展。

第三，由于本次采集的语料时间跨年度为 2019 年 1 月 29 日至 2020 年 11 月 4 日，而在 2020 年初又暴发了严重的新冠肺炎疫情，所以在天山网新疆新闻词表中，可以发现一些有关疫情的词的频率较高。例如，词频排名第三的“新增”，例如包含它的新闻标题有：

例 13.4 月 22 日新疆（含兵团）无新增新冠肺炎确诊病例（2020 年 4 月 23 日）

例 14. 喀什目前所有无症状感染者未明确有疑似病例、确诊病例、发热病人接触史（2020 年 10 月 26 日）

例 15. 新疆（含兵团）新增 1 例新型冠状病毒感染的肺炎确诊病例（2020 年 2 月 1 日）

另外，从与疫情有关的新闻标题中，可以反映出新疆维吾尔自治区各级党委、政府、社会群体对新冠疫情的重视以及正确应对。例如：

例 16. 自治区召开视频会议研究部署新型冠状病毒感染的肺炎疫情防控工作（2020 年 1 月 28 日）

例 17. 心理专家谈疫情：要有节制地获取信息（2020 年 1 月 29 日）

例 18.【众志成城 打赢疫情防控阻击战】新疆 13 家信息技术企业在防疫期间显身手（2020 年 2 月 21 日）

我们要进一步提高政治站位，坚持把疫情防控作为重大政治责任、摆在突出位置，深入贯彻落实习近平总书记关于做好常态化疫情防控工作的重要指示精神，坚决打好疫情防控阻击战，坚决维护各族群众生命安全和身体健康。

第四，全疆 1660000 平方千米，拥有 2523.22 万人口（截至 2019 年年末截止），新疆维吾尔自治区实现地区生产总值（GDP）13797.58 亿元（截至 2020 年），从各个方面来看，新疆都具有极其重要的战略地位。在表 1 中，许多高频词条也体现了新疆对脱贫攻坚、经济发展等方面的重视程度。例如，“企业”“项目”“服务”“建设”“启动”“工作”“中国”“脱贫”“就业”等词条。包含有这些词条的新闻标题都可以展现出新疆社会稳定形势发生根本变化，各族人民群众的获得感、幸福感、安全感显著增强。新疆一直把经济发展放在重要位置，要以推进丝绸之路经济带核心区建设为引领，推进新疆贸易持续高质量发展。例如：

例 19. 新疆投入 1 亿专项资金支持中小企业发展（2019 年 10 月 11 日）

例 20. 乌鲁木齐 2020 年将高标准建设河马泉新区 同时加快两河片区基础设施建设（2020 年 1 月 18 日）

例 21. 阿巴・阿尤甫的脱贫色彩：从沙漠黄到辣椒素红（2020 年 7 月 1 日）

以上高频词条以及新闻标题，反映了新疆各级党委、政府、社会各级，对保障社会持续稳定、推动经济平稳发展、不断改善营商环境、不断改善人民生活环境、三大攻坚战取得重大进展、民族团结、宗教和谐等方面做出了重大贡献。新疆深入贯彻落实习近平总书记重要讲话指示精神和党中央决策部署，经济社会发展和民生改善取得了前所未有的成就，脱贫攻坚取得了决定性成就。新疆媒体将以充沛饱满的热情，讲述好脱贫攻坚故事，弘扬好脱贫攻坚精神，分享好脱贫攻坚经验，为推动新疆经济社会发展贡献力量。

结语

本文借助 Python 进行语料收集，使用 NLPIR-ICTCLAS 汉语分词系统进行中文词切分，使用 MyZiciFreq 字词频率统计工具进行词频统计，建立了天山网新疆新闻标题词表。之后，对具有社会性、地域性的高频词进行提取和分析研究。所以，新闻标题语料库通过高频词汇，可以反映出该新闻媒体的侧重点，也可以反映出新闻编者经常使用简略的高频重点词汇来体现出正文的内容。但是，本文所建立的语料库还有原始语料不足、时间跨度太小等问题，因此没有对天山网新疆新闻标题进行全面、详细、深入的描写。另外，新疆新闻媒体还有许多，新闻也有标题、正文等多个方面急需研究，本文仅以天山网新疆新闻的标题作为研究对象，研究范围还是稍显狭小。在新疆，除了网络新闻媒体，报纸刊物还有新疆日报、乌鲁木齐晚报、兵团日报、阿克苏日报等，它们都十分缺乏研究。综上，对新疆新闻媒体，今后还需要时间跨度足够大、原始语料足够多、研究内容足够深入、研究范围足够广泛的研究。媒

参考文献

[1] 马子恩. 热点事件新闻语料库的研制及词汇研究 [D]. 南京：南京师范大学，2012.

[2] 班文冲 . 基于语料库的网络新闻标题词频研究 [D]. 曲阜：曲阜师范大学文学院，2016.

（下转第64页）